

7. Databases of Genome and pathway networks

Benjamin F. Matthews

United States Department of Agriculture
Soybean Genomics and Improvement
Laboratory

Beltsville, MD 20708

matthewb@ba.ars.usda.gov

What we will cover today

- Identifying the biological function and context of a gene
- KEGG metabolic database
- Pathway analysis
- Integrate and interpret molecular work within biological context
- <http://www.genome.ad.jp/kegg/>
- Single nucleotide polymorphism (SNP) identification



KEGG: Kyoto Encyclopedia of Genes and Genomes

- Metabolic Database
- Linked to other databases
 - DBGET (database retrieval system)
 - <http://www.genome.jp/dbget/>
 - GenomeNet
- Flat file data
- GIF images
- Java graphics
- 3D images

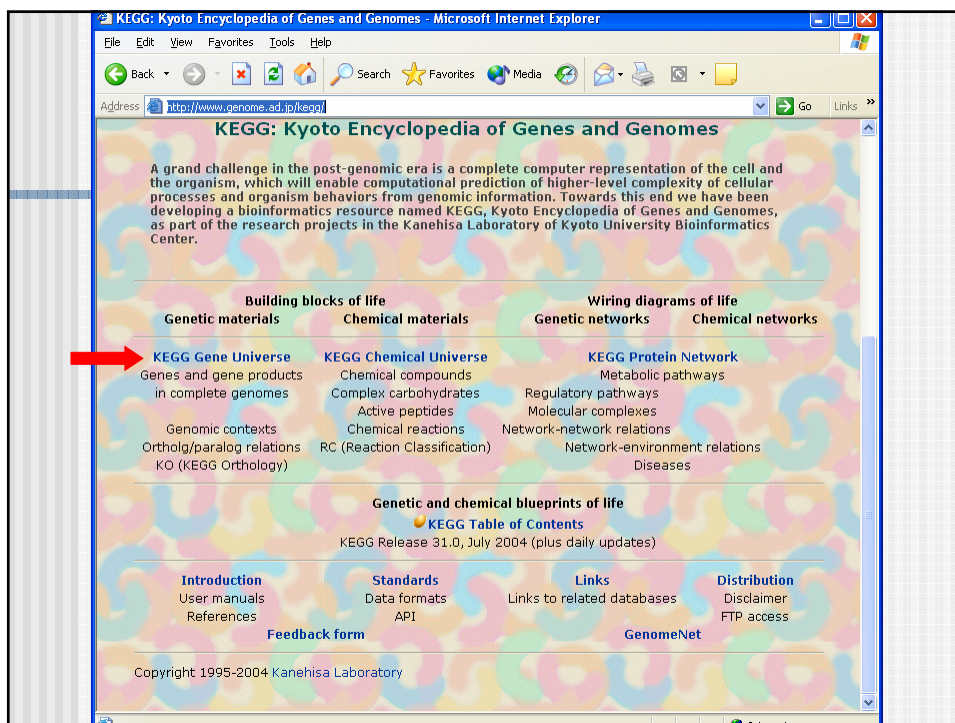
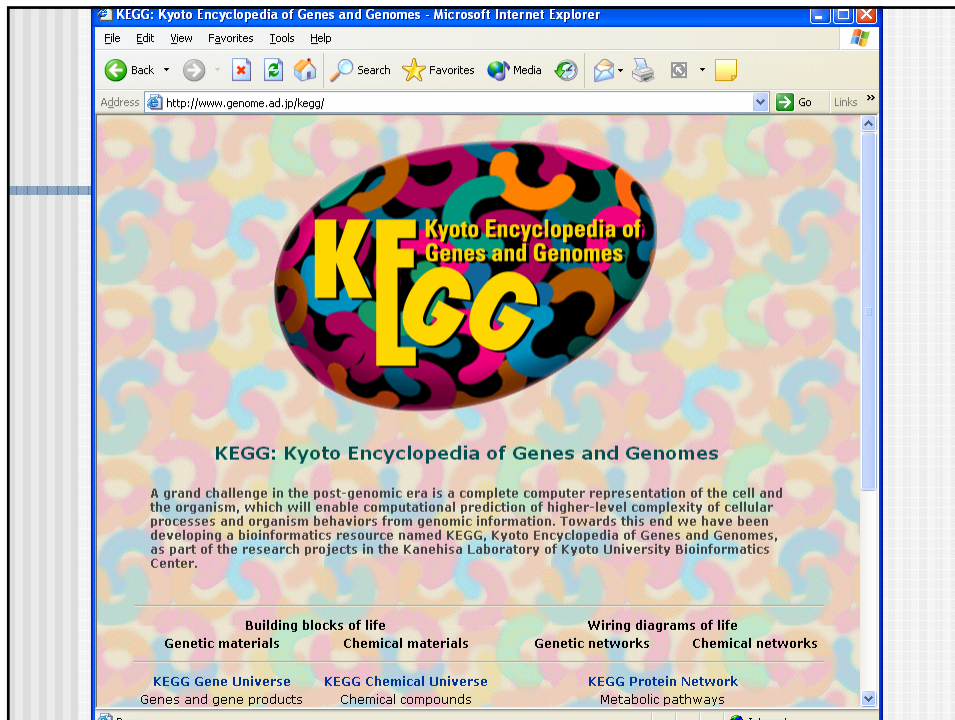
<http://www.genome.ad.jp/kegg/>

KEGG: Kyoto Encyclopedia of Genes and Genomes

- Linked to sequence interpretation tools
 - BLAST
 - FASTA
 - MOTIF
 - CLUSTALW

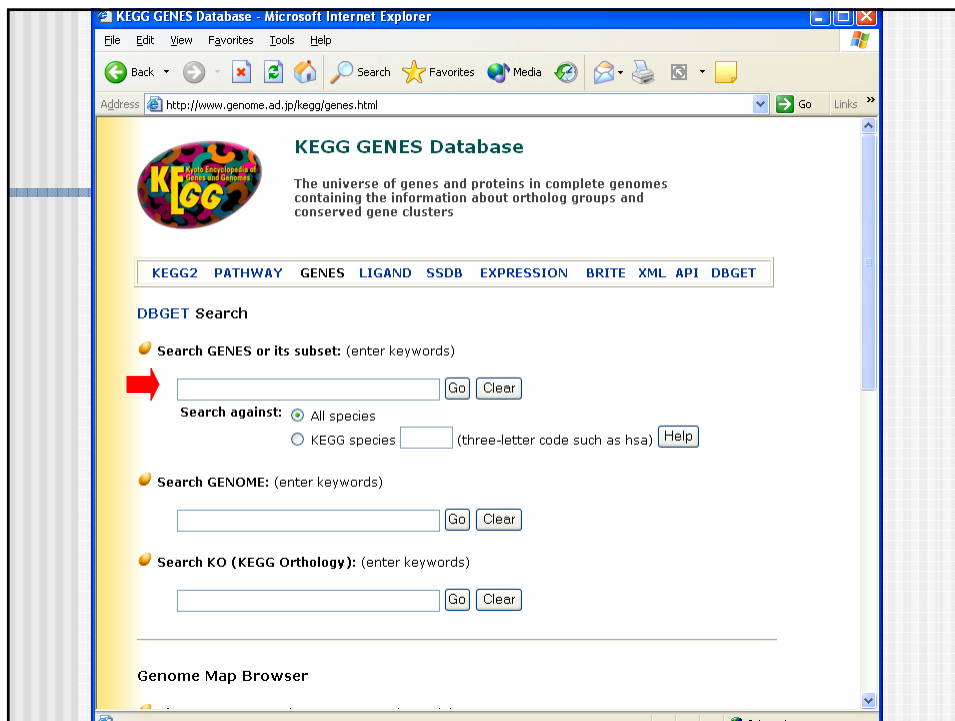
What is the function of a gene or gene product?

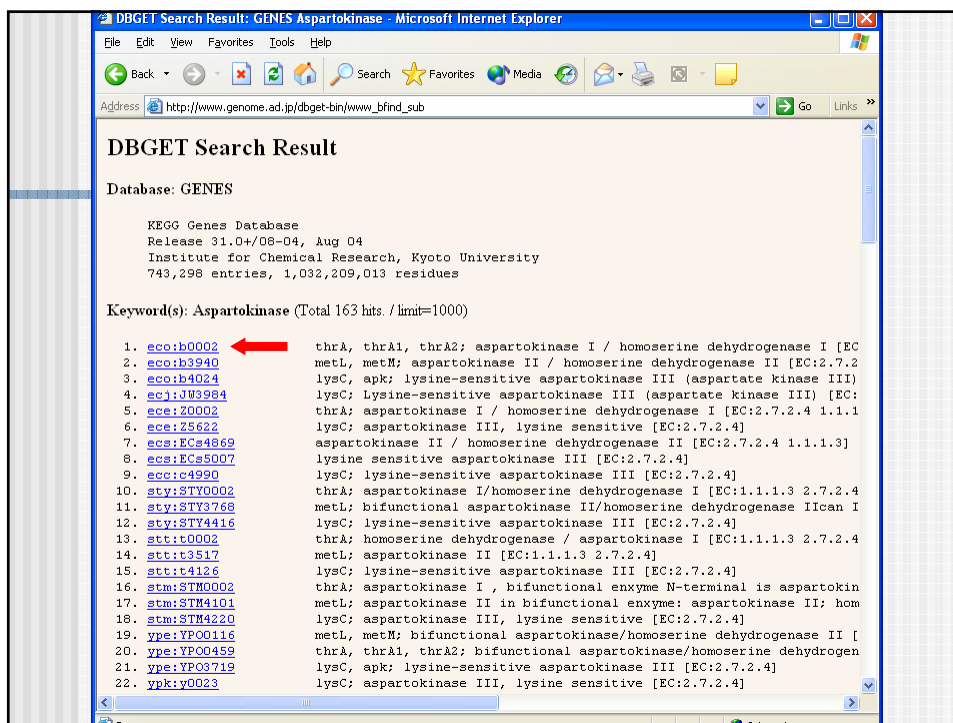
What is its biological Context?



KEGG: Kyoto Encyclopedia of Genes and Genomes

- Gene Universe
 - Genes and gene products
 - Genomic contexts
 - Ortholog/paralog relations
- Chemical Universe
 - Chemical compounds
 - Peptides, carbohydrates
 - Chemical reactions
- Protein network
- Metabolic pathways
- Regulatory pathways
- Networks and complexes





DBGET Result: E.coli b0002 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Print Mail

Address http://www.genome.ad.jp/dbget-bin/www_bget?eco:b0002

K166 Escherichia coli K-12 MG1655: b0002

Help

Entry	b0002 CDS E.coli
Gene name	thrA, thrA1, thrA2
Definition	aspartokinase I / homoserine dehydrogenase I [EC:2.7.2.4 1.1.1.3]
K0	K0: K00003 homoserine dehydrogenase K0: K00928 aspartate kinase OC search OC viewer
Class	Metabolism; Amino Acid Metabolism; Glycine, serine and threonine metabolism [PATH:eco00260] Metabolism; Amino Acid Metabolism; Lysine biosynthesis [PATH:eco00300] Gene catalog
SSDB	Ortholog Paralog Motif Gene cluster
Other DBs	Wisconsin: b0002 Colibri: thrA RegulonDB: ECK120000002 NCBI: 1786183 UniProt: P00561
LinkDB	PDB All DBs
Position	337..2799 Genome map
AA seq	820 aa AA seq FASTA-genes FASTA-sp BLAST-nr MRVLKFGGTSVANAERFLRVAD ILESNARQGVATVLSAPAKITNHLVAMIEKTISGQDA LPNISDAERIFAELLTGLAAQPGFPLAQKTFVDQEFQAKHVLHGISLLGQCPSINA ALICRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIP ADHMVLMAGFTAGNEKSELVVLGRNGSDYSAAVLAACLRADCEIWTVDVGVYTCDFPRQV PDAELLKSMSTQAEMLSYFGAKVLHPRTITP IAOQFIPCLIKNTGNPQAPGLIGASRD EDELVPKGISLNNMAMFVSVPGMKGMVGMARVFAAMSRARISVVLITOSSEYSISF CVPQSDCVRAERAMQEEFYLELEKGLLEPLAVTERLAIISVVGDGMRTLRGISAKFFAAL ARANINIVAI AQSSERSISVYVNNDDATTGVRVTHOMLFNTDQVIEVPVIGVGVGGAL LEQLKROQSWLKNKHIDLRVCGVANSKALLTNVHGLNLENAQELAQAKFPNGLIRL VKEYHLLNPVIVDCTSSQAVADQYADFLREGFHVYTPNKKANTSEMDYYHQLRYAAEKSR RKFLYDTNVSAGLPVIEHLQMLNAGDELMKFSGLSGSLSYIFGKLDGMSFSEATTLA REMGYTEFPDDLSGMDVARKLLILARETGRELEADIEIEPVLPAEFNAEGVAAFPMA NLSQDLDDLFAARVAKARDEKGLVRYGNIEDGVCVRVIAEVDGNDPLFKVKNGENALAF TSHYQPLFLVLRGTAGNDVTAAGVFADLLRLTSLMKLGV

Internet

DBGET Result: E.coli b0002 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Print Mail

Address http://www.genome.ad.jp/dbget-bin/www_bget?eco:b0002

Position 337..2799 [Genome map](#)

AA seq 820 aa [AA seq](#) [FASTA-genes](#) [FASTA-sp](#) [BLAST-nr](#)

MRVLKFGGTSVANAERFLRVAD ILESNARQGVATVLSAPAKITNHLVAMIEKTISGQDA
LPNISDAERIFAELLTGLAAQPGFPLAQKTFVDQEFQAKHVLHGISLLGQCPSINA
ALICRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIP
ADHMVLMAGFTAGNEKSELVVLGRNGSDYSAAVLAACLRADCEIWTVDVGVYTCDFPRQV
PDAELLKSMSTQAEMLSYFGAKVLHPRTITP IAOQFIPCLIKNTGNPQAPGLIGASRD
EDELVPKGISLNNMAMFVSVPGMKGMVGMARVFAAMSRARISVVLITOSSEYSISF
CVPQSDCVRAERAMQEEFYLELEKGLLEPLAVTERLAIISVVGDGMRTLRGISAKFFAAL
ARANINIVAI AQSSERSISVYVNNDDATTGVRVTHOMLFNTDQVIEVPVIGVGVGGAL
LEQLKROQSWLKNKHIDLRVCGVANSKALLTNVHGLNLENAQELAQAKFPNGLIRL
VKEYHLLNPVIVDCTSSQAVADQYADFLREGFHVYTPNKKANTSEMDYYHQLRYAAEKSR
RKFLYDTNVSAGLPVIEHLQMLNAGDELMKFSGLSGSLSYIFGKLDGMSFSEATTLA
REMGYTEFPDDLSGMDVARKLLILARETGRELEADIEIEPVLPAEFNAEGVAAFPMA
NLSQDLDDLFAARVAKARDEKGLVRYGNIEDGVCVRVIAEVDGNDPLFKVKNGENALAF
TSHYQPLFLVLRGTAGNDVTAAGVFADLLRLTSLMKLGV

NT seq 2463 nt [NT seq](#) +upstream 0 nt +downstream 0 nt

atcgagagtgttgaaagtcggcggtacatcagtgaggcaaatgcagaacgtttctgcgtgtt
gccgatattctgaaagaacatgccaggcagggcaggtggccaccgtctctctgcccc
gccaaaatcaccacaccctggtggcgatgattgaaaaaacattagcggccaggtgct
ttaccgaatcagcgatgccgaacgtatttttgcggaacttttgacgggactgcgcc
gccagcgggggttcccgctggcgcaattgaaactttctgcgacaggaattggccaa
ataaaacatgctctgcattgctgctggtggcgagtgccggatagcatcaacgct
gcgctgatttgcgctggcgagaaatgctgacgcatcattatggcggcggtattagaagc
cgcggtcacacgttactgttatcgatccggcgaataactgctggcagtgggcattac
ctcgaaatcaccgtcgatattgctgagtcaccccgctatttgcggcaagccgcatccg
gctgacacatggtgctgattggcaggtttcccgccggttaataaaaaaggcgaactggtg
gtgcttgagcgaacggttccgactactctgctggcggtgctggctgctgtttacgcgc
gattgttgcgagatttggacggagcttgacggggctctacactgcgaccccgctcaggtg
ccgcatgcgaggtgttgaagtcgattgctaccaggaagcagtgagctttctacttc
ggcgctaaagtcttccaccccgaccattaccccatcgccaggttcagatcccttgc
ctdattaaaaatccgaaatctctcaaacaccagctacatctatctccacacatgatt

Done Internet

DBGET Result: E.coli b0002 - Microsoft Internet Explorer

Address: http://www.genome.ad.jp/dbget-bin/www_bget?eco:b0002

KEGG Escherichia coli K-12 MG1655: b0002

Help

Entry	b0002 CDS E.coli
Gene name	thrA, thrA1, thrA2
Definition	aspartokinase I / homoserine dehydrogenase I [EC:2.7.2.4 1.1.1.3]
KO	KO: K00003 homoserine dehydrogenase KO: K00928 aspartate kinase OC search OC viewer
Class	Metabolism; Amino Acid Metabolism; Glycine, serine and threonine metabolism [PATH:eco00260] Metabolism; Amino Acid Metabolism; Lysine biosynthesis [PATH:eco00300] Gene catalog
SSDB	Ortholog Paralog Motif Gene cluster
Other DBs	Wisconsin: b0002 Colibri: thrA RegulonDB: ECK12000002 NCBI: 1786183 UniProt: P00561
LinkDB	PDB All DBs
Position	337..2799 Genome map
AA seq	820 aa AA seq FASTA-genes FASTA-sp BLAST-nr MRVLKFGGTSVANAERFLRVADILESNAHQGVATVLSAPAKITNHLVAMIEKTISGQDA LPNISDAERIFAEELLTGLAAQPGFPLAQKTFVDQEFQIKHVLHGISLLGQCPDSINA ALICRSEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIP ADHMVLMAGPTAGWEKSELVVLGRNGSDYSAAVLAACLRADECCEIWTVDVGVTCDPRQV

Search Result: b0002 (Escherichia coli K-12 Genes) - Microsoft Internet Explorer

Address: http://www.genome.ad.jp/dbget-bin/find_www_sub?text=E.coli.kegg&keywords=b0002&option=a

Escherichia coli K-12 Genes

According to the KEGG pathways

5. Amino Acid Metabolism

5.3 Glycine, serine and threonine metabolism [PATH:eco00260]

*b0002 thrA, thrA1, thrA2; aspartokinase I / homoserine dehydrogenase I [EC:2.7.2.4 1.1.1.3] [SP:AKIH_ECOLI]

*b0002 thrA, thrA1, thrA2; aspartokinase I / homoserine dehydrogenase I [EC:2.7.2.4 1.1.1.3] [SP:AKIH_ECOLI]

5.8 Lysine biosynthesis [PATH:eco00300]

*b0002 thrA, thrA1, thrA2; aspartokinase I / homoserine dehydrogenase I [EC:2.7.2.4 1.1.1.3] [SP:AKIH_ECOLI]

*b0002 thrA, thrA1, thrA2; aspartokinase I / homoserine dehydrogenase I [EC:2.7.2.4 1.1.1.3] [SP:AKIH_ECOLI]

Last updated: Aug 04, 2004

[KEGG | DBGET | GenomeNet]

Search Result: 5.3 Glycine, serine and threonine metabolism (Escherichia coli K-12 Genes) - Microsoft Internet Explorer

Address: http://www.genome.ad.jp/dbget-bin/hfind_www_sub?foldflag=T&foldopt=w&localfiledr=files&htext=E.coli.kegg&keywords=5.3+Glycine,+sei

Escherichia coli K-12 Genes

According to the KEGG pathways

5. Amino Acid Metabolism

5.3 Glycine, serine and threonine metabolism [PATH:eco00260]

- *b0002 thrA, thrA1, thrA2; aspartokinase I / homoserine dehydrogenase I [EC:2.7.2.4 1.1.1.3] [SP:AKIH_ECOLI]
- *b3940 metL, metM; aspartokinase II / homoserine dehydrogenase II [EC:2.7.2.4 1.1.1.3] [SP:AK2H_ECOLI]
- *b4024 lysC, apk; lysine-sensitive aspartokinase III (aspartate kinase III) [EC:2.7.2.4] [SP:AK3_ECOLI]
- *b3433 asd, hom; aspartate-semialdehyde dehydrogenase (Asa dehydrogenase) [EC:1.2.1.11] [SP:DHAS_ECOLI]
- *b0002 thrA, thrA1, thrA2; aspartokinase I / homoserine dehydrogenase I [EC:2.7.2.4 1.1.1.3] [SP:AKIH_ECOLI]
- *b3940 metL, metM; aspartokinase II / homoserine dehydrogenase II [EC:2.7.2.4 1.1.1.3] [SP:AK2H_ECOLI]
- b0003 thrB; homoserine kinase (HK) [EC:2.7.1.39] [SP:KHSE_ECOLI]
- *b0004 thrC; threonine synthase [EC:4.2.3.1] [SP:THRC_ECOLI]
- b0870 ybJ; L-allo-threonine aldolase (L-allo-TA) (L-allo-threonine acetaldehyde-lyase) [EC:4.1.2.5] [SP:LTAE_ECOLI]
- *b2551 glyA; serine hydroxymethyltransferase (serine methylase) (SHMT) [EC:2.1.2.1] [SP:GLYA_ECOLI]
- b1849 purT; phosphoribosylglycinamide formyltransferase 2 (part 2) (GAR transferase 2) (5'-phosphoribosylglycinamide transferase 2) (formate-dependent GAR transferase) [EC:2.1.2.-] [SP:FURT_ECOLI]
- b4388 serB; phosphoserine phosphatase (PSP) (0-phosphoserine phosphohydrolase) [EC:3.1.3.3] [SP:SERB_ECOLI]
- *b0907 serC, pdcF; phosphoserine aminotransferase [EC:2.6.1.52] [SP:SEPC_ECOLI]
- b2913 serA; D-3-phosphoglycerate dehydrogenase (PGDH) [EC:1.1.1.95] [SP:SERA_ECOLI]
- *b0514 ybbZ; hypothetical 38.7 kD protein in GIP-fdrA intergenic region [EC:2.7.1.31] [SP:GRK1_ECOLI]
- *b3124 yhaD; glycerate kinase 2 [EC:2.7.1.31] [SP:GRK2_ECOLI]

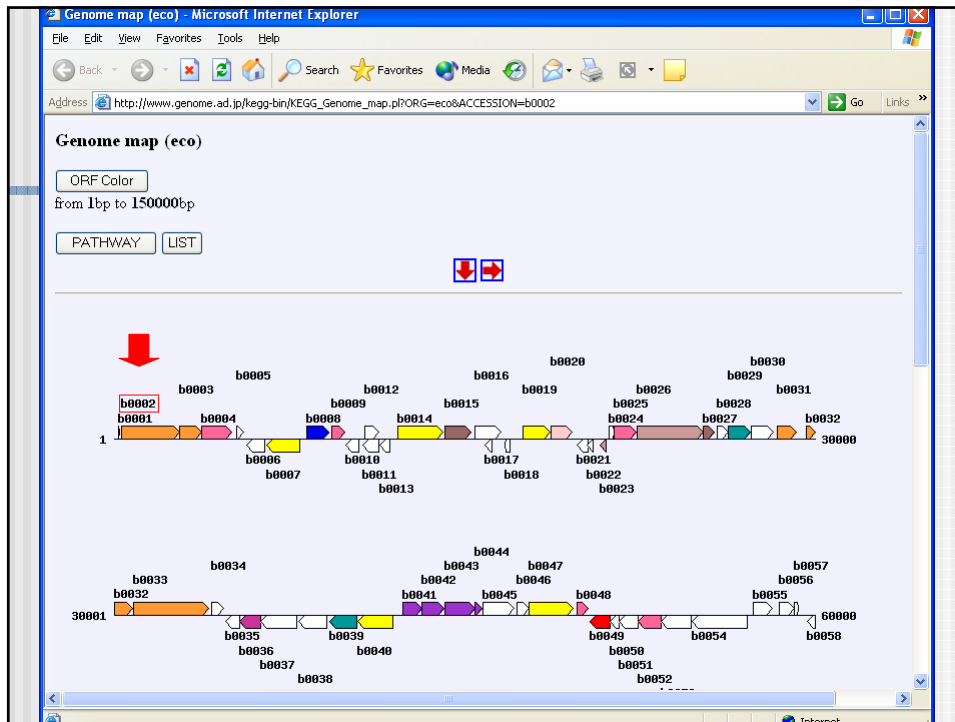
DBGET Result: E.coli b0002 - Microsoft Internet Explorer

Address: http://www.genome.ad.jp/dbget-bin/www_bget?eco:b0002

Escherichia coli K-12 MG1655: b0002

Help

Entry	b0002 CDS E.coli
Gene name	thrA, thrA1, thrA2
Definition	aspartokinase I / homoserine dehydrogenase I [EC:2.7.2.4 1.1.1.3]
KO	KO: K00003 homoserine dehydrogenase KO: K00928 aspartate kinase OC search OC viewer
Class	Metabolism; Amino Acid Metabolism; Glycine, serine and threonine metabolism [PATH:eco00260] Metabolism; Amino Acid Metabolism; Lysine biosynthesis [PATH:eco00300] Gene catalog
SSDB	Ortholog Paralog Motif Gene cluster
Other DBs	Wisconsin: b0002 Colibri: thrA RegulonDB: ECK120000002 NCBI: 1786183 UniProt: P00561
LinkDB	PDB All DBs
Position	337..2799 Genome map
AA seq	820 aa AA seq FASTA-genes FASTA-sp BLAST-nr MRVLKFGGTSVANAERFLRVADILESNAQQGVATVLSAPAKITNHLVAMIEKTISGQDA LPNISDAERIFAELLTGLAAAPGFPLAQKTFVDQEFQIKHVLHGISLLGQCPSINA ALICRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIP ADHMFLMAGFTAGNEKGEVLVLRNGSDYSAAVLAAACLRADCEIWDVDGVYTCDEPROV



KEGG: Kyoto Encyclopedia of Genes and Genomes - Microsoft Internet Explorer

Address: <http://www.genome.ad.jp/kegg/>

KEGG: Kyoto Encyclopedia of Genes and Genomes


A grand challenge in the post-genomic era is a complete computer representation of the cell and the organism, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic information. Towards this end we have been developing a bioinformatics resource named KEGG, Kyoto Encyclopedia of Genes and Genomes, as part of the research projects in the Kanehisa Laboratory of Kyoto University Bioinformatics Center.

Building blocks of life		Wiring diagrams of life	
Genetic materials	Chemical materials	Genetic networks	Chemical networks
KEGG Gene Universe Genes and gene products in complete genomes Genomic contexts Ortholog/paralog relations	KEGG Chemical Universe Chemical compounds Complex carbohydrates Active peptides Chemical reactions RC (Reaction Classification)	KEGG Protein Network Metabolic pathways Regulatory pathways Molecular complexes Network-network relations Network-environment relations	

KEGG PATHWAY Database - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.genome.ad.jp/kegg/pathway.html> Go Links



KEGG PATHWAY Database

Current knowledge on molecular interaction networks, including metabolic pathways, regulatory pathways, and molecular complexes

KEGG2 PATHWAY GENES LIGAND SSDB EXPRESSION BRITE XML API DBGET

Go to:

- 1. Metabolism**
Carbohydrate Energy Lipid Nucleotide Amino acid Other amino acid
Glycan PK/NRP Cofactor/vitamin Secondary metabolite Xenobiotics
- 2. Genetic Information Processing**
- 3. Environmental Information Processing**
- 4. Cellular Processes**
- 5. Human Diseases**

See also: KO (KEGG Orthology)

1. Metabolism

1.1 Carbohydrate Metabolism

Glycolysis / Gluconeogenesis	Ortholog, Oxidoreductases
Citrate cycle (TCA cycle)	Ortholog
Pentose phosphate pathway	Ortholog
Pentose and glucuronate interconversions	Ortholog
Fructose and mannose metabolism	Ortholog, PTS
Galactose metabolism	Ortholog
Ascorbate and aldarate metabolism	Ortholog
Starch and sucrose metabolism	Ortholog, PTS
Aminosugars metabolism	Ortholog, PTS
Nucleotide sugars metabolism	Ortholog

KEGG PATHWAY Database - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.genome.ad.jp/kegg/pathway.html> Go Links

1.4 Nucleotide Metabolism

Purine metabolism	Ortholog
Pyrimidine metabolism	Ortholog

1.5 Amino Acid Metabolism

Glutamate metabolism	Ortholog, Aminotransferases
Alanine and aspartate metabolism	Ortholog
Glycine, serine and threonine metabolism	Ortholog
Methionine metabolism	Ortholog
Cysteine metabolism	Ortholog
Valine, leucine and isoleucine degradation	Ortholog
Valine, leucine and isoleucine biosynthesis	Ortholog, PTS
Lysine biosynthesis	Ortholog
Lysine degradation	
Arginine and proline metabolism	Ortholog
Histidine metabolism	Ortholog
Tyrosine metabolism	Ortholog
Phenylalanine metabolism	Ortholog
Tryptophan metabolism	
Phenylalanine, tyrosine and tryptophan biosynthesis	Ortholog
Urea cycle and metabolism of amino groups	Ortholog

1.6 Metabolism of Other Amino Acids

beta-Alanine metabolism	Ortholog
Taurine and hypotaurine metabolism	
Aminophosphonate metabolism	
Selenoamino acid metabolism	
Cyanoamino acid metabolism	
D-Glutamine and D-glutamate metabolism	
D-Arginine and D-ornithine metabolism	
D-Alanine metabolism	
Glutathione metabolism	

1.7 Glycan Biosynthesis and Metabolism

N-Glycans biosynthesis	
N-Glycan degradation	

DBGET Result: A.thaliana At1g31230 At4g19710 At5g13280 At5g14060 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.genome.ad.jp/dbget-bin/www_bget?ath+At1g31230+At4g19710+At5g13280+At5g14060

KEGG Arabidopsis thaliana: At1g31230

Help

Entry	At1g31230 CDS A.thaliana
Gene name	F28K20.19
Definition	bifunctional aspartate kinase/homoserine dehydrogenase / AK-HSDH [EC:2.7.2.4 1.1.1.3]
KO	KO: K00003 homoserine dehydrogenase KO: K00928 aspartate kinase OC search OC viewer
Class	Metabolism; Amino Acid Metabolism; Glycine, serine and threonine metabolism [PATH:ath00260] Metabolism; Amino Acid Metabolism; Lysine biosynthesis [PATH:ath00300] Gene catalog
SSDB	Ortholog Paralog Motif Gene cluster
Other DBs	TIGR: At1g31230 TAIR: At1g31230 MIPS: At1g31230 NCBI: 15221653
LinkDB	PDB All DBs
Position	1: complement(join(11158725..11158970,11159067..11159159,11159234..11159317,11159402..11159537,11159642..11159787,11159921..11160146,11160227..11160309,11160395..11160589,11160712..11160810,11160936..11161049,11161138..11161225,11161305..11161513,11161603..11161655,11161771..11162074,11162179..11162267,11162383..11162757,11162841.. 11163036)) Genome map
AA seq	911 aa AA seq FASTA-genes FASTA-sp BLAST-nr

DBGET Result: A.thaliana At1g31230 At4g19710 At5g13280 At5g14060 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.genome.ad.jp/dbget-bin/www_bget?ath+At1g31230+At4g19710+At5g13280+At5g14060

KEGG Arabidopsis thaliana: At4g19710

Help

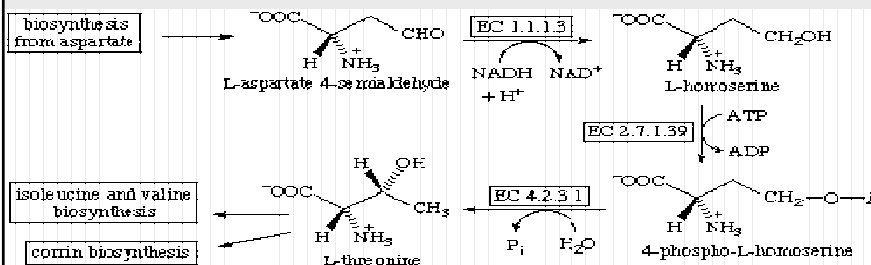
Entry	At4g19710 CDS A.thaliana
Gene name	T16H5.70
Definition	bifunctional aspartate kinase/homoserine dehydrogenase, putative / AK-HSDH, putative [EC:1.1.1.3 2.7.2.4]
KO	KO: K00003 homoserine dehydrogenase KO: K00928 aspartate kinase OC search OC viewer
Class	Metabolism; Amino Acid Metabolism; Glycine, serine and threonine metabolism [PATH:ath00260] Metabolism; Amino Acid Metabolism; Lysine biosynthesis [PATH:ath00300] Gene catalog
SSDB	Ortholog Paralog Motif Gene cluster
Other DBs	TIGR: At4g19710 TAIR: At4g19710 MIPS: At4g19710 NCBI: 42572959
LinkDB	PDB All DBs
Position	IV: join(10725239..10725443,10725521..10725901,10725975..10726063,10726147..10726450,10726517..10726569,10726648..10726856,10726965..10727052,10727144..10727257,10727386..10727484,10727597..10727791,10727898..10727980,10728135..10728360,10728483..10728628,10728721..10728856,10728952..10729035,10729121..10729213,10729301..10729546) Genome map
AA seq	916 aa AA seq FASTA-genes FASTA-sp BLAST-nr

Enzyme Nomenclature

- <http://www.chem.qmul.ac.uk/iubmb/enzyme/>
- Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB)
- Reactions and pathways

Threonine Biosynthesis

When cursor points to a box further details will be displayed in the status window below. If you click on the box, you will change to appropriate reaction scheme or enzyme specification.



Return to:

[enzymes](#) homepage.

[aspartate](#) biosynthesis.

[isoleucine and valine](#) biosynthesis.

[corrin](#) biosynthesis.

[EC 1.1.1.3](#) homoserine dehydrogenase

[EC 2.7.1.39](#) homoserine kinase

[EC 4.2.3.1](#) threonine synthase

Single Nucleotide Polymorphism (SNP) Discovery

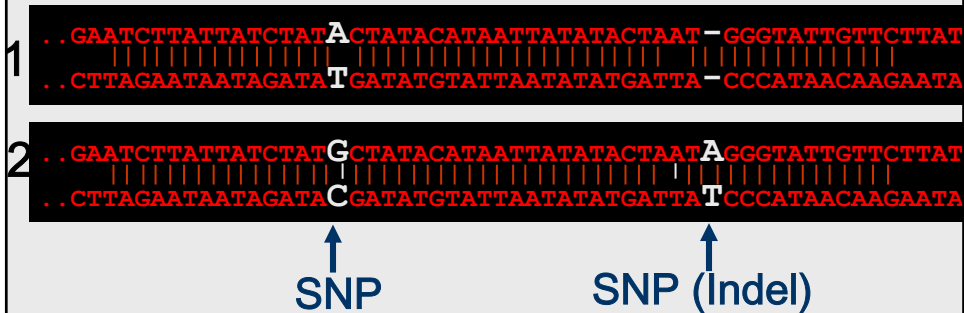
Use of SNPs?

- Genetic mapping
 - Highly abundant markers
 - Easy to assay
- Distinguish plant variety
 - Variety protection
- Phylogenetics

Single Nucleotide Polymorphism

- A working definition -

- **Single base changes between homologous DNA fragments**
- **Small insertions and deletions (indels)**



SNPs (Single Nucleotide Polymorphisms)

- Genetic variation
- Can be alleles of genes
- also differences in non-coding regions collected from genome sequencing of different individuals
- **dbSNP** at the NCBI - all public SNP data
- **SNP Consortium** at CSHL - high quality set

SNP Marker Discovery

1. Design PCR primers to existing sequence: Complete genes, ESTs, BAC-ends, BAC-subclones, SSR flanking regions, etc.
2. Identify sequence tagged sites (STSs) visually (agarose gel) and by sequence analysis
3. Amplify the homologous sequence from a panel of diverse genotypes
4. Determine sequence quality (PHRED) and align sequence traces (PHRAP) from the diverse genotypes
5. Analyze assemblies with SNP discovery software such as PolyBayes for SNP discovery in redundant sequence
6. Analyze haplotype variation and database

Discovery of SNPs in aligned DNA sequence data using PolyBayes in the Consed viewer.

File Navigate Info Color Dim Misc Help

010831.nap.ace.1 Contig1 Sone Tags Pos: clear

Search for String Compl Cont Compare Cont Find Main Win Err/10kb: 286.19

90 100 110 120 130 140 150 160 170

CONSensus GACTCATTACCGTTGGATCATCATGAAATTGTGCATCAAGTTCGGAACCTATTCCAAACATTTTCACCGTTGGAATTTACGAAGA

F02_010831Minsoy_.g GACTCATTACCGTTGGATCATCATGAAATTGTGCATCAAGTTCGGAACCTATTCCAAACATTTTCACCGTTGGAATTTACGAAGA

F01_010831Archer_.g GACTCATTACCGTTGGATCATCATGAAATTGTGCATCAAGTTCGGAACCTATTCCAAACATTTTCACCGTTGGAATTTACGAAGA

F05_010831PI20933.g GACTCATTACCGTTGGATCATCATGAAATTGTGCATCAAGTTCGGAACCTATTCCAAACATTTTCACCGTTGGAATTTACGAAGA

F04_010831Evans_.g GACTCATTACCGTTGGATCATCATGAAATTGTGCATCAAGTTCGGAACCTATTCCAAACATTTTCACCGTTGGAATTTACGAAGA

F06_010831Peking_.g GACTCATTACCGTTGGATCATCATGAAATTGTGCATCAAGTTCGGAACCTATTCCAAACATTTTCACCGTTGGAATTTACGAAGA

F03_010831Noir_1_.g GACTCATTACCGTTGGATCATCATGAAATTGTGCATCAAGTTCGGAACCTATTCCAAACATTTTCACCGTTGGAATTTACGAAGA

F04_010831Evans_.b GACTCATTACCGTTGGATCATCATGAAATTGTGCATCAAGTTCGGAACCTATTCCAAACATTTTCACCGTTGGAATTTACGAAGA

F06_010831Peking_.b GACTCATTACCGTTGGATCATCATGAAATTGTGCATCAAGTTCGGAACCTATTCCAAACATTTTCACCGTTGGAATTTACGAAGA

F03_010831Noir_1_.b GACTCATTACCGTTGGATCATCATGAAATTGTGCATCAAGTTCGGAACCTATTCCAAACATTTTCACCGTTGGAATTTACGAAGA

F02_010831Minsoy_.b GACTCATTACCGTTGGATCATCATGAAATTGTGCATCAAGTTCGGAACCTATTCCAAACATTTTCACCGTTGGAATTTACGAAGA

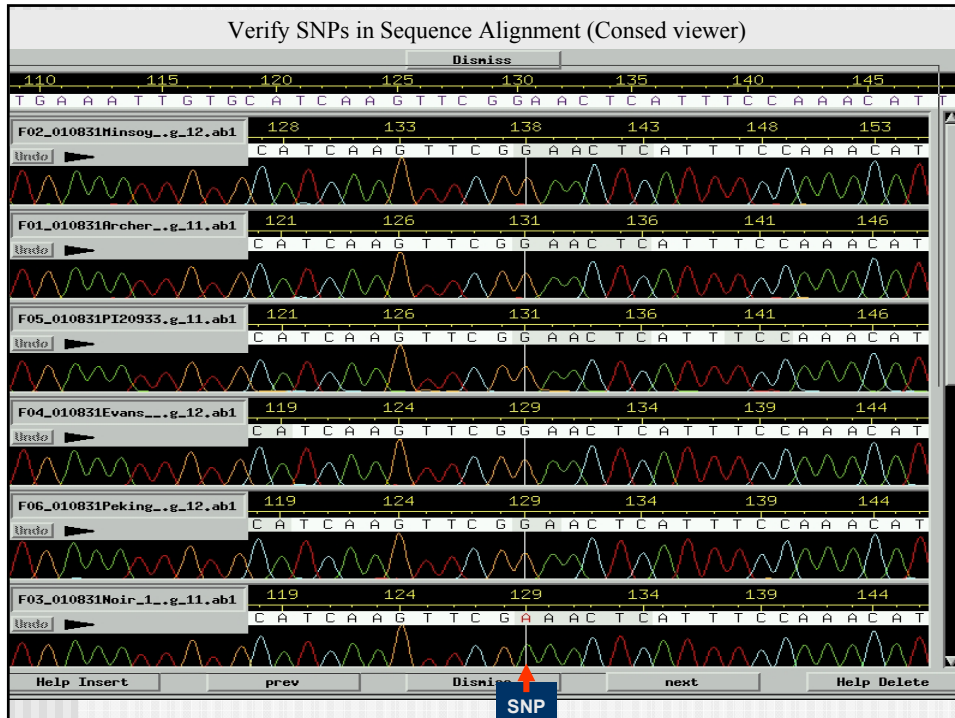
F05_010831PI20933.b GACTCATTACCGTTGGATCATCATGAAATTGTGCATCAAGTTCGGAACCTATTCCAAACATTTTCACCGTTGGAATTTACGAAGA

F01_010831Archer_.b GACTCATTACCGTTGGATCATCATGAAATTGTGCATCAAGTTCGGAACCTATTCCAAACATTTTCACCGTTGGAATTTACGAAGA

PolyBayes SNP Discovery software

SNP

<<< << < Prev Next > >> >>> cursor dismiss



Data from Consed to SNP Database

1. Consensus sequence is created from the Phrap alignment after trimming of bases with Phred score < 25 from each end of sequence.
2. The position of each single base change or insertion/deletion (indel) is indicated in the consensus sequence e.g. Single base changes: A/T or G/A and indels: T/- or TCGG/-
3. The "haplotypes" that are present in the fragment is determined and the haplotype of each genotype is determined and placed in the database. The haplotype is the linear, ordered arrangement of SNP alleles on a chromosome or DNA fragment.
4. Sequence data are placed in the database in a format that allows direct submission to dbSNP, the National Center for Biotechnology Information database for SNP data from all species.

Relational Database Structure

1. Source File:

- Information relating to the source of the sequence
- BARC sequence ID
- PCR primer sequences
- optimal PCR conditions
- person submitting information

2. SNP File:

- Data for each individual SNP
- BARC sequence ID
- consensus sequence derived from the alignment of alternative genotypes,
- sequence to the 5' side of the SNP
- the alternative SNP alleles
- the sequence to the 3' side of the SNP

The Source File

Submitter Name	Youlin	Date entered into database	
Sequence ID	7548		
Name			
Source ID	AB003680		
GB Accession #	AB003680		7/31/2002
Sequence Source	Genbank		
Type	gene		
Putative Phenotype Association	random		
Comment	A3B4 Glycinin		
Forward Primer	AGTGCCCAATATGTTGTCCTCTAC		
Reverse Primer	GTTGCGCTTCAAGTTCCAAT		
Forward Pr ID	7548		
Reverse Pr ID	7548		
Mg Conc in mM	3.75		
BUFFER:	STB-S		
Annealing Temp	50		
Primer Plate			
Forward primer location			
Reverse primer location			
Forward Pr Conc in uM	0.15		
Reverse Pr Conc in uM	0.15		
PCR Product Length	551		
Seq Data Length	3358		
Comments @ PCR Amplif			
Comments @ sequencing			
Source sequence	AGTGCCCAATATGTTGTCCTCTACAGGGTATGTAATTCATTTCATATACAAAGTAATCAACATGAACTAATATACGTGCATACCTGCCATCTACCATAGTAGTGTTTTGTGGATTTTCAGTGTTAATTAGTGATCTTCAGAGAAAGAAATAAAGAAAGCACTAAACAGGGGGAAATCATAATTCATAGGTCATATACCATACAATAAGAAGACATAAAATGTTAACAAGTATGTTGTAGGGTTGGGTTCTTTAATGTCATTTAAATTAATCTCACTTTGATAGATAACTGATTTTATAGAGGTTATGTAGAGGTAATTTTATAGTTATAATGGAGTAAATGTTTGTAATCTAAATTTGTGCATTGATTTTTTAAAGTGAGTTTCCACATAT		

The SNP File

SNP Name	BARC-GM-00001	Date entered into Database	7/31/2002
Sequence ID	7548	Date submitted to NCBI	5/31/2002
SNP Position in consensus	363	Genebank ss#	4473759
SNP position in consed		Submitter name	Youlin
Probability	1		
Trans type	S		
GB Accession #	AB003680	Sequence Source	Genbank
Type	gene	Putative Phenotype Association	random
Comment	A3B4 Glycinin		
5' side	ATTACAAGTAATCAACATGAACTAATATACGTGCATACTTGACATCTACCATA GTAGTGTTTTGTGGATTTTCAGTGTTAATTAGTGTATCTTCAGAGAAAGAAATAA AAGAAAGCACTAAAAGAGGGGGAAATCATAATTCATAGGTCATATACGATACAA TAAGAAGACATAAAATGTTAACAAGTATGTTGTAGGGTTGGGTTCCCTTTAATG TCATTTAAATTAAATCTCACTTTGATAGATAACTGATTTTTAGAGGTTATGTAGAG GTAATTTTATAGTTATAATGGAATAAAATTGTTTGATTCTAAATTTGTGCATTGAT TTTTTAAAGTGAGTTTCGACATATAATT		
SNP	T/C		
3' side	AAAATATATCATTACCTCTTATTTGATAATAATTAAACATTTATCATTATATAATAA TAATAATAATATGTAACATGTATTATTATATCCATGCATGCAGAA		

Relational Database Structure

3. Genotype File:

- BARC sequence ID
- SNP name
- the SNP position in the consensus sequence
- SNP allele present in each genotype

4. Haplotype File: (One record for each haplotype for a sequence ID)

- BARC sequence ID
- haplotype ID
- SNP position(s) in the consensus sequence
- the genotype(s) with the haplotype
- the complete sequence of the haplotype

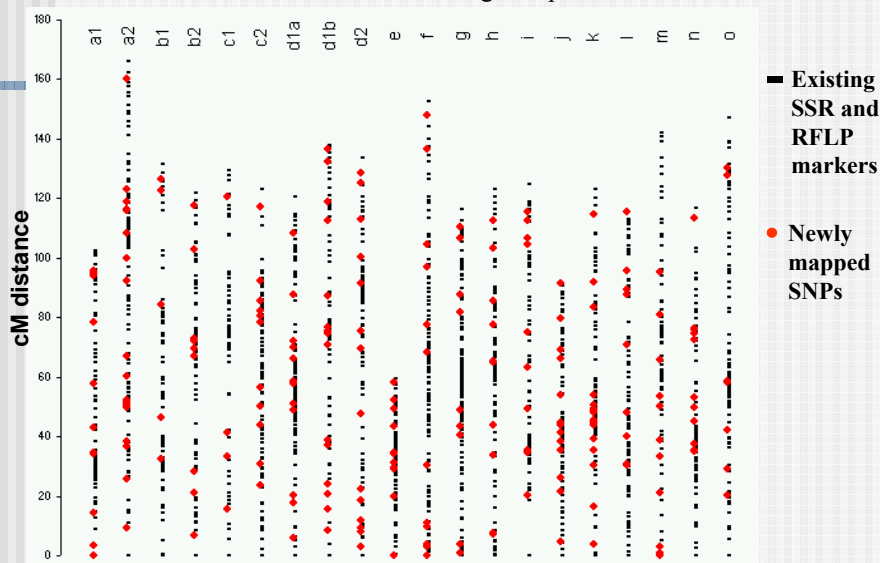
The Haplotype File

Sequence ID	14843	Date entered into Database	1/3/2003
Haplotype ID	014843_HAP2	Submitter Name	Ik-Young
SNP Positions	449-476-		
Haplotype Short Seq	[ta]		
Member Genotypes	Archer_1;Minsoy_;		
Haplotype Long Seq	GATACCCTAGATGAAACAGCAAGGAGTGCAGGAAGCACACGCTCCACAAGGTCA CCCAATACAAGAAGGGCAAGGACAGCATCGCCGCTCAGGGAAAAACGCCGTTAT GACCGCAAAACAGTCCGGTTACGGTGGCCAGACCAAGCCCGTTTTCCACAAAAA GGTACAATTCTACTGTTTCCCCAAACACTGCCTTATGGTTTTATTGCTAATAAT TTACTTTTAATTAATTAATTAGGCGAAAAACCACCAAGAAAATTGTGTTGAGGCTCCA GTGCCAAGGATGCAAGCATGTCTCGCAGCACGCTATCAAGGTAAGCATCACTT TCCGTCGTCGTTTTGTAITGTTAATTGTGCATGTTTGGTTTACAGTCGAAAATGCTG CTGTGCACGAACCAATGCAAAATCCAATGAATAAAAGTAAATGAACGCAGAAGGT TGGCAAAATTACTTTTGCCTCAAACAGTAATTTTGCAACTGATTCCCAACATGCT CATTGCCGTTTTGAACACGTGGTTGATGCTTGCTATGTGGTTCTGTTTTGCAGAG		
Consensus Sequence	GATACCCTAGATGAAACAGCAAGGAGTGCAGGAAGCACACGCTCCACAAGGTCA CCCAATACAAGAAGGGCAAGGACAGCATCGCCGCTCAGGGAAAAACGCCGTTAT GACCGCAAAACAGTCCGGTTACGGTGGCCAGACCAAGCCCGTTTTCCACAAAAA GGTACAATTCTACTGTTTCCCCAAACACTGCCTTATGGTTTTATTGCTAATAAT TTACTTTTAATTAATTAATTAGGCGAAAAACCACCAAGAAAATTGTGTTGAGGCTCCA GTGCCAAGGATGCAAGCATGTCTCGCAGCACGCTATCAAGGTAAGCATCACTT TCCGTCGTCGTTTTGTAITGTTAATTGTGCATGTTTGGTTTACAGTCGAAAATGCTG CTGTGCACGAACCAATGCAAAATCCAATGAATAAAAGTAAATGAACGCAGAAGGT TGGCAAAATCACTTTTGCCTCAAACAGTAATTTTGCAACTGATTCCCAACATGCT CATTGCCGTTTTGAACACGTGGTTGATGCTTGCTATGTGGTTCTGTTTTGCAGAG GTGCAAGCACTTTGAGATCGGTGGTGACAAGGAAGGAAAAAGAACATCTCTCTT CTAGATGGAAATATTAGAATTTGCAATTCACACACTCTCTGTACTTTTCTCAGTTT GATTGCCACGGG		

The Genotype File

Sequence ID	7548	Date entered into Database	7/31/2002
SNP name	BARC-GM-00001	Submitter name	Youlin
SNP Position	363		
SNP position before trimming			
Archer	T		
Minsoy	T		
Noir 1	T		
Evans	T		
P1209332	C		
Peking	T		
Poly McN	1		
Poly McA	1		
Poly 200xEvans	0		
Poly six genotypes	0		

Positions of 455 SNPs Mapped on the 20 Linkage Groups of the Soybean Genetic Linkage Map



What we covered today

- KEGG Metabolic database
- Pathway analysis
- Integrate and interpret molecular work within biological context
- <http://www.genome.ad.jp/kegg/>
- Single nucleotide polymorphism (SNP) identification